A GENERAL SOLUTION TO THE MAXIMIZATION OF THE MULTIDIMENSIONAL GENERALIZED RAYLEIGH QUOTIENT USED IN LINEAR DISCRIMINANT ANALYSIS FOR SIGNAL CLASSIFICATION

Ramon E. Prieto

Electrical Engineering Department,
Stanford University
rprieto@stanford.edu

ABSTRACT

A more general solution and a new didactic demonstration of the maximization of the multidimensional case of the Generalized Rayleigh Quotient are described. This solution will not only be the well-known eigenvectors solution widely available in the literature but also a general transformation that is not necessarily orthogonal. The demonstration uses only basic linear algebra and simple lagrangian maximization to find the transformation matrix that maximizes the multidimensional Generalized Rayleigh Quotient for Linear Discriminant Analysis, widely used in signal classification applications.

1. INTRODUCTION

Linear discriminant analysis (LDA) have been widely used as a signal classification tool in many signal processing applications such as speech recognition and image processing.

In speech recognition applications such as one described in [1], LDA is applied to the mel-cepstral feature space, linearly transforming the mel-cepstral coefficients such that less parameters are used by the hidden markov models in the pattern recognition stage. This parameters, at the same time, will have most of the discrimination information between the different phones or triphones given the chosen dimensionality.

In image processing applications such as the face recognition method described in [2] they use LDA to reduce the number of features after projecting the original space using non-linear kernel mapping. Specially in this case, where the dimensionality of an image is high, LDA provides a way to reduce the computational cost by reducing the number of dimensions while keeping high discrimination between classes.

As seen in both of the examples above, the aim of LDA is to reduce the dimensionality of the analysis space such that the discrimination between different signal classes is maximum given a certain measure. The measure that is used to quantify the discrimination between different signal classes using only one dimension is the Generalized Rayleigh Quotient, that is defined as ([3]):

$$Q(w) = \frac{w^T S_B w}{w^T S_W w} \qquad w \in \mathbb{R}_n, \qquad S_W \in \mathbb{R}_{n\times n},\ S_B \in \mathbb{R}_{n\times n} \tag{1}$$

If we want to measure the discrimination between signal classes in a higher dimensional space, we can use the d-dimensional version of the quotient that is:

$$Q(W) = \frac{|W^T S_B W|}{|W^T S_W W|} \qquad W \in \mathbb{R}_{n\times d}, \qquad S_W \in \mathbb{R}_{n\times n},\ S_B \in \mathbb{R}_{n\times n} \tag{2}$$

where S_B is the between class scatter matrix of the training data, it's rank is $r \geq d$ and it is symmetric and positive semi-definite (PSD). S_W is the within class scatter matrix of the training data, it is full rank and positive definite (PD). W is a full rank linear transformation that will transform the original feature space into a lower dimensional space of size d.

The objective of this paper is to to find a more general definition of the argument that maximizes $Q(W)$, showing that there is an infinite number of transformations W that maximize $Q(W)$ and formally defining it's characteristics. As we can see in formula 65 we will find that the eigenvectors solution widely available in literature ([3]) is not a general solution.

At the same time we show a demonstration, different than the ones available in the literature, that uses only basic linear algebra and lagrangian maximization to find the general solution of the problem of maximizing $Q(W)$. This, at the same time, will serve as a new proof for the specific case when the maximum of $Q(W)$, in formula 2, is attained when the columns w_i of W are eigenvectors such that $S_B w_i = \lambda_i S_W w_i$.

For this, we first define the quotient maximization problem and it's widely known eigenvectors solution in section 2. To reach our general solution, in section 3, we show a demonstration in in 5 different steps that can be grouped in the following way:

1. Reduce the Quotient maximization problem to a determinant maximization problem in sections 3.1 and 3.2.
2. In the determinant maximization problem, bound the determinant, maximize the bound and show that, for the argument that maximizes the bound, both the determinant and the bound are equal. (sections 3.3 and 3.4).
3. Recover the original solution from the solution of the determinant maximization problem. Show that, the eignevectors solution is not the only solution and there is an infinite number of solutions to the problem. (section 3.5)

2. PROBLEM

The maximization problem is stated as:

$$\widehat{W} = arg \max_{W \in \mathbb{R}_{n \times d}} (Q(W)) \tag{3}$$

$$= arg \max_{W \in \mathbb{R}_{n \times d}} \left(\frac{|W^T S_B W|}{|W^T S_W W|} \right) \tag{4}$$

where:

$$W \in \mathbb{R}_{n \times d},\ full\, rank.$$
$$S_W \in \mathbb{R}_{n \times n},\ S_B \in \mathbb{R}_{n \times n}.$$
$$S_W\ is\ PD,\ S_B\ is\ PSD.$$
$$The\ rank\ of\ S_B\ is\ r \geq d$$

The widely known solution to this problem is such that ([3]):

$$S_B w_i = \lambda_i S_W w_i \qquad w_i \in \mathbb{R}_n, 0 \leq i \leq d \tag{5}$$
$$W = [w_1\ w_2\ \ ...\ w_d]$$

According to this specific solution, the multidimensional Generalized Rayleigh quotient is maximized when the columns of W are the d eigenvectors of $S_W^{-1} S_B$ corresponding to the d highest eigenvalues.

This problem was initially analyzed by Fisher ([4]), and it's solution gave birth to the method of statistical analysis called linear discriminant analysis (LDA). A more precise treatment of the problem is given in [5] where they analyze independently the unidimensional case and the multidimensional case of the problem. In the demonstration of the multidimensional case, the solution is found by maximizing several unidimensional cases, one for each column of W. As a result, the non general eigenvectors solution is found.

We shall now proceed with a demonstration that will allow us to find a more general solution.

3. DEMONSTRATION

3.1. Decompose S_W into it's cholesky components

Since S_W is PSD we can decompose it using it's cholesky factorization:

$$S_W = RR^T \qquad R\ is\ lower\ triangular,\ R \in \mathbb{R}_{n \times n} \tag{6}$$
$$\left|W^T S_W W\right| = \left|W^T R R^T W\right| \tag{7}$$

then, we can define Y as the transformation of W by R^T and replace it into formula 2:

$$Y = R^T W, \quad W = R^{-T} Y \tag{8}$$
$$Q(W) = \frac{|Y^T R^{-1} S_B R^{-T} Y|}{|Y^T Y|} \tag{9}$$

Given that W is full rank and Sw is PD, then Y is full rank.

3.2. Decompose S_W into it's SVD components

Given that Y is full rank, then it's singular value decomposition (SVD, [6]) is such that:

$$Y = U \Sigma V^T \qquad U \in \mathbb{R}_{n \times d}, \Sigma \in \mathbb{R}_{d \times d}, V \in \mathbb{R}_{d \times d}. \tag{10}$$
$$W = R^{-T} U \Sigma V^T \tag{11}$$

where the columns of U are orthonormal, the columns of V are orthonormal and the matrix Σ is a diagonal matrix containing the singular values of Y: σ_i. Then, we can reduce our problem even further realizing that:

$$Y^T Y = V \Sigma U^T U \Sigma V^T \tag{12}$$
$$Y^T Y = V \Sigma^2 V^T \tag{13}$$
$$\left|Y^T Y\right| = \prod_{i=1}^{d} \sigma_i^2 \tag{14}$$

and applying it to formula 9:

$$Q(W) = \frac{|Y^T R^{-1} S_B R^{-T} Y|}{\prod_{i=1}^{d} \sigma_i^2} \tag{15}$$
$$= \frac{|V \Sigma U^T R^{-1} S_B R^{-T} U \Sigma V^T|}{\prod_{i=1}^{d} \sigma_i^2} \tag{16}$$
$$= \frac{|V|\,|\Sigma|\,|U^T R^{-1} S_B R^{-T} U|\,|\Sigma|\,|V^T|}{\prod_{i=1}^{d} \sigma_i^2} \tag{17}$$
$$= \frac{|U^T R^{-1} S_B R^{-T} U| \prod_{i=1}^{d} \sigma_i^2}{\prod_{i=1}^{d} \sigma_i^2} \tag{18}$$
$$= \left|U^T R^{-1} S_B R^{-T} U\right| \tag{19}$$

For convenience in the following steps of our demonstration, we now define the matrix M and the function $P(U)$, and replace them into formula 19:

$$M = R^{-1} S_B R^{-T} \qquad M \in \mathbb{R}_{n \times n} \tag{20}$$
$$P(U) = Q(W) = \left|U^T M U\right| \tag{21}$$

From formula 20 we can see that M has the same rank as S_B (r), it is symmetric and PSD.

As we see in formula 21, we have reduced $Q(W)$ to a single determinant. Then our problem is transformed into:

$$\widehat{U} = arg \max_{U \in \mathbb{R}_{n \times d}} P(U) = arg \max_{U \in \mathbb{R}_{n \times d}} \left|U^T M U\right| \tag{22}$$
$$The\ columns\ of\ U\ are\ orthonormal.$$

We can already conclude that the solution of our original problem is independent of V and Σ. As it will become evident in section 3.5, the variation of V and Σ is the difference between our general solution and the restricted eigenvectors solution.

3.3. Bounding the determinant

We know from [6] that the determinant of a PSD matrix is less or equal than the product of it's diagonal elements. Taking this bound into account, representing M by it's eigenvalue decomposition (M has rank r) and applying it to formula 21:

$$M = \sum_{i=1}^{r} \lambda_{M_i} q_{M_i} q_{M_i}^T \tag{23}$$

$$Z = U^T M U \tag{24}$$

$$= \sum_{i=1}^{r} \lambda_{M_i} U^T q_{M_i} q_{M_i}^T U \tag{25}$$

$$Z_{jj} = \sum_{i=1}^{r} \lambda_{M_i} \left(u_j^T q_{M_i} \right)^2 \tag{26}$$

$$|Z| = P(U) = \left| U^T M U \right| \tag{27}$$

$$\leq \prod_{j=1}^{d} Z_{jj} = \prod_{j=1}^{d} \sum_{i=1}^{r} \lambda_{M_i} \left(u_j^T q_{M_i} \right)^2 \tag{28}$$

where Z_{jj} are the diagonal elements of Z, u_j are the column vectors of U, and q_{M_i} and λ_{M_i} are the eigenvectors and eigenvalues (ordered in descending order, $\lambda_{M_i} \geq \lambda_{M_{i+1}} > 0$) of M respectively. As a result, formula 28 sets an upper bound for the determinant stated in formula 21.

3.4. Maximizing the Bound and Maximizing the determinant

From formula 28 we can make the following statement: If we maximize the bound such that u_j are orthonormal, and the argument $\widehat{u_j}$ found makes $|\widehat{Z}| = \prod_{j=1}^{d} \widehat{Z}_{jj}$, then we can say that:

$$|Z| \leq \prod_{j=1}^{d} Z_{jj} \leq \prod_{j=1}^{d} \widehat{Z}_{jj} = \left| \widehat{Z} \right| \tag{29}$$

$$|Z| \leq \left| \widehat{Z} \right| \tag{30}$$

and this means that, we would have found the matrix $\widehat{U}$ that maximizes $|Z| = P(U)$ and we would have solved the problem stated in formula 22.

Following this reasoning, in this section, we will find a matrix $\widehat{U}$ that is the argument that maximizes $\prod_{j=1}^{d} Z_{jj}$. Then we will prove that the matrix $\widehat{U}$ found is such that there is an equality relationship between $P(\widehat{U}) = |\widehat{Z}|$ and $\prod_{j=1}^{d} \widehat{Z}_{jj}$. Then we will conclude that $\widehat{U}$ maximizes $P(U)$.

Let's first define $J(U)$ as the bound found in formula 28:

$$J(U) = \prod_{j=1}^{d} Z_{jj} = \prod_{j=1}^{d} \sum_{i=1}^{r} \lambda_{M_i} \left(u_j^T q_{M_i} \right)^2 \tag{31}$$

Let's then solve the problem of maximizing the bound $J(U)$:

$$\widehat{U} = arg \max_{\|u_j\|=1} J(U) \tag{32}$$

Given that M is PSD, then the values λ_{M_i} are greater than zero for $1 \leq i \leq r$ and $J(U)$ is convex over every u_j. Having $J(U)$ as a convex function guarantees that applying a lagrangian maximization method on it will give a global maximum. Then, we now define the lagrangian $L(U)$ and we proceed to maximize it:

$$L(U) = J(U) + \sum_{l=1}^{d} \lambda_l \left(1 - \|u_l\|^2 \right) \tag{33}$$

$$\frac{\partial L\left(\widehat{U}\right)}{\partial \widehat{u}_k} = 2J\left(\widehat{U}\right) \frac{\sum_{i=1}^{r} \lambda_{M_i} \left(\widehat{u}_k^T q_{M_i}\right) q_{M_i}}{\sum_{i=1}^{r} \lambda_{M_i} \left(\widehat{u}_k^T q_{M_i}\right)^2} - 2\lambda_k \widehat{u}_k \tag{34}$$

$$= 0 \tag{35}$$

$$\lambda_k \widehat{u}_k = J\left(\widehat{U}\right) \frac{\sum_{i=1}^{r} \lambda_{M_i} \left(\widehat{u}_k^T q_{M_i}\right) q_{M_i}}{\sum_{i=1}^{r} \lambda_{M_i} \left(\widehat{u}_k^T q_{M_i}\right)^2} \tag{36}$$

We now left-multiply both sides of formula 36 by $\widehat{u}_k^T$ and rewrite it again. As a result we get the solution $\widehat{u}_k$:

$$J\left(\widehat{U}\right) \frac{\sum_{i=1}^{r} \lambda_{M_i} \left(\widehat{u}_k^T q_{M_i}\right)^2}{\sum_{i=1}^{r} \lambda_{M_i} \left(\widehat{u}_k^T q_{M_i}\right)^2} = \lambda_k \widehat{u}_k^T \widehat{u}_k = \lambda_k \tag{37}$$

$$J(\widehat{U}) = \lambda_k \tag{38}$$

$$\frac{\sum_{i=1}^{r} \lambda_{M_i} \left(\widehat{u}_k^T q_{M_i}\right) q_{M_i}}{\sum_{i=1}^{r} \lambda_{M_i} \left(\widehat{u}_k^T q_{M_i}\right)^2} = \widehat{u}_k \tag{39}$$

To simplify our answer, we left-multiplying both sides of formula 39 by $q_{M_h}^T$:

$$\frac{\sum_{i=1}^{r} \lambda_{M_i} \left(\widehat{u}_k^T q_{M_i}\right) q_{M_h}^T q_{M_i}}{\sum_{i=1}^{r} \lambda_{M_i} \left(\widehat{u}_k^T q_{M_i}\right)^2} = q_{M_h}^T \widehat{u}_k \tag{40}$$

$$\frac{\lambda_{M_h} \left(\widehat{u}_k^T q_{M_h}\right) q_{M_h}^T q_{M_h}}{\sum_{i=1}^{r} \lambda_{M_i} \left(\widehat{u}_k^T q_{M_i}\right)^2} = q_{M_h}^T \widehat{u}_k \tag{41}$$

$$\frac{\lambda_{M_h} \widehat{u}_k^T q_{M_h}}{\sum_{i=1}^{r} \lambda_{M_i} \left(\widehat{u}_k^T q_{M_i}\right)^2} = q_{M_h}^T \widehat{u}_k \tag{42}$$

Since every λ_{M_i} is greater than zero, then, formula 42 is true for every k and every h if and only if:

$$For\ every\ integer\ k \in [1\ d]\ there\ is\ an\ integer \tag{43}$$

$$h \in [1\ r]\ such\ that \tag{44}$$

$$\widehat{u}_k^T q_{M_h} = 1 \tag{45}$$

$$\widehat{u}_k^T q_{M_l} = 0\ for\ l \neq h \tag{46}$$

To maximize $J(U)$ we will choose h such that q_{M_h} correspond to the highest eigenvalues λ_{M_h}, i.e. we will chose $k = h$:

$$\widehat{u}_k^T q_{M_k} = 1 \tag{47}$$

$$\widehat{u}_k^T q_{M_l} = 0,\ for\ l \neq k \tag{48}$$

Adding the fact that $\|\widehat{u}_k\| = 1$:

$$\widehat{u}_k = q_{M_k} \tag{49}$$

In formula 49 we have just found the solution of the problem of maximizing the upper bound of the determinant defined in formula 21, that is that the eigenvectors of the matrix M maximize the bound $J(U)$.

Now we will show that $P(\widehat{U}) = J(\widehat{U})$ (i.e. the determinant of $|\widehat{U}^T M \widehat{U}|$ is equal to the product of it's diagonal components). For this, we first notice that the resulting $\widehat{Z}$ is a diagonal matrix:

$$\widehat{Z} = \widehat{U}^T M \widehat{U} \tag{50}$$

$$= \Lambda_{M_d} \tag{51}$$

where Λ_{M_d} contains the highest d eigenvalues of M. Given that $\widehat{Z}$ is a diagonal matrix, then, it's determinant is equal to the product of it's diagonal components:

$$P(\widehat{U}) = \left|\widehat{Z}\right| = \prod_{i=1}^{d} \widehat{Z}_{ii} = J(\widehat{U}) \tag{52}$$

Concluding with this section of the demonstration, we have found $\widehat{U}$ with orthonormal columns such that, for every matrix $U \in \mathbb{R}_{n\times d}$ that has orthonormal columns, the following is true:

$$P(U) = \left|U^T M U\right| \leq J(U) \leq J(\widehat{U}) = P(\widehat{U}) \tag{53}$$

that is the same as saying that $\widehat{U}$ (i.e. the matrix which columns are the eigenvectors of M) maximizes $P(U)$.

3.5. Transforming the solution into one for the original problem

To transform the solution we found for the determinant maximization problem of formula 22 into a solution for the original problem of formula 4 we first rewrite formula 49:

$$M\widehat{U} = \widehat{U}\Lambda_{M_d} \tag{54}$$

$$\tag{55}$$

and combining this with formulas 10, 8, 20 and 6:

$$MU\Sigma V^T = \widehat{U}\Lambda_{M_d}\Sigma V^T \tag{56}$$

$$M\widehat{Y} = \widehat{U}\Sigma V^T V\Lambda_{M_d}V^T \tag{57}$$

$$M\widehat{Y} = \widehat{Y}V\Lambda_{M_d}V^T \tag{58}$$

$$MR^T\widehat{W} = R^T\widehat{W}V\Lambda_{M_d}V^T \tag{59}$$

$$R^{-1}S_B R^{-T}R^T\widehat{W} = R^T\widehat{W}V\Lambda_{M_d}V^T \tag{60}$$

$$S_B\widehat{W} = RR^T\widehat{W}V\Lambda_{M_d}V^T \tag{61}$$

$$S_B\widehat{W} = S_W\widehat{W}V\Lambda_{M_d}V^T \tag{62}$$

Formula 62 states an implicit version of the general solution for the problem in formula 4 where, as we have seen in formulas 17 and 18, the matrix V can be chosen arbitrarily as far as it's columns are orthonormal.

If we choose V to be the identity matrix, then the solution of the original problem for that specific case is:

$$S_B\widehat{W} = S_W\widehat{W}I\Lambda_{M_d}I \tag{63}$$

$$= S_W\widehat{W}\Lambda_{M_d} \tag{64}$$

what means that there is a specific solution that maximizes the multidimensional Generalized Rayleigh Quotient, and that specific solution is a matrix which columns are the eigenvectors of the system S_B, S_W. This is the well known eigenvectors solution.

We shall now write an explicit formula for the definition of the general solution of maximization of the multidimensional Generalized Rayleigh Quotient (formula 4). That solution is built based on formulas 49 and 11:

$$\widehat{W} = R^{-T}\widehat{U}\Sigma V^T \tag{65}$$

$$R\ is\ the\ cholesky\ decomposition\ of\ S_W$$

$$\Sigma\ is\ any\ diagonal\ matrix\ ,\ \Sigma \in \mathbb{R}_{d\times d}$$

$$V\ is\ any\ orthogonal\ matrix\ ,\ V \in \mathbb{R}_{d\times d}$$

$$R^{-1}S_B R^{-T}\widehat{U} = \widehat{U}\Lambda_{M_d}$$

4. CONCLUSIONS

We have found a more general solution to the maximization of the multidimensional generalized rayleigh quotient. Our solution doesn't only include the well known eigenvectors solution, but also an infinite number of solutions where our projection matrix can be non orthogonal. To find this solution we have divided the demonstration in 5 parts, where we use cholesky decomposition and singular value decomposition to reduce the problem to a determinant maximization problem. Then we have used a property of the determinant of a PSD matrix to find a bound for that determinant. Then we find the argument that maximizes that bound, and we show that argument maximizes the determinant too. At the end, given the solution of the determinant maximization problem we recover the solution of the original problem. Both implicitly and explicitly.

5. REFERENCES

[1] Hunt, M., and Lefbvre, C., "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech", in Proc. ICASSP'89, Glasgow, UK, May 1989, pp 262-265.

[2] Liu Q., Huang R., Lu H., Ma S., "Face Recognition Using Kernel Based Fisher Discriminant Analysis", in Proc. 5th IEEE Inter. Conf. on Automatic Face and Gesture Recognition, May 2002, pp 187 -191.

[3] Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification, A Wiley-Interscience Publication, Second Edition, 2001.

[4] Fisher, R.A., The statistical utilization of multiple measurements, Eugen., Vol. 8, pp. 376-386, 1938.

[5] Wilks, S.S., Mathematical Statistics, New York: Wiley, 1962.

[6] B. Noble, J.W. Daniel, Applied Linear Algebra, Prentice-Hall Inc, 1977.